Haozhu Wang

Cluster Report on patient Data using Weka

I have used the Windows version of Weka on Windows Vista, using the file “train-a.csv” with JVM memory heap allocated at 1536MB. For the preprocessing part, I have removed the much of the attributes like “Record ID”, all of “Other Dx codes”, “Other PR codes”, etc. narrowing down the 79 attributes to 34 attributes. Those ID are removed because they are either irrelevant or I don’t know what they mean.

First time Run

The first time I ran clustering algorithm on Weka, I just use SimpleKMeans, test mode on traning set data(the first option). The scheme uses Euclidean Distance, maximum of 500 iterations, and total of 6 clusters. This is just a standard procedure. Below are the screenshots of summary output of the cluster run. I used screenshot because I am on a windows machine and can’t redirect the output to file.

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -N 6 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10
Relation:     train-a-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.Remove-R55-69-weka.filters.
Instances:    99999
Attributes:   34
              PatientID
              Record Count
              Interval
              Patient-Sex
              Age
              Patient-Race
              Patient-Ethnicity
              Patient-Disposition
              Length of Stay
              Admit-Type
              Admit-Source
              Hospital-ID
              Region-ID
              Principal-Dx-Code
              Admit-Dx-Code
              Principal-PR-Code
              Cause-E-Code
              Place-E-Code
              Reimb DRG
              Reimb MDC
              AccomCharges
              AncilCharges
              TotalCharges
              Serv-Class
              Emergency-Dept-Ind
              Pot Amb
              Complication-Minor
              Complication-Sever
              Trauma-Minor
              Trauma-Severe
              Trauma-Severity
              Nosocomial Inf
              Severity
              Cost Weight
Test mode:    evaluate on training data

=== Model and evaluation on training set ===
```

kMeans
======

Number of iterations: 49
Within cluster sum of squared errors: 175394.64933301846
Missing values globally replaced with mean/mode

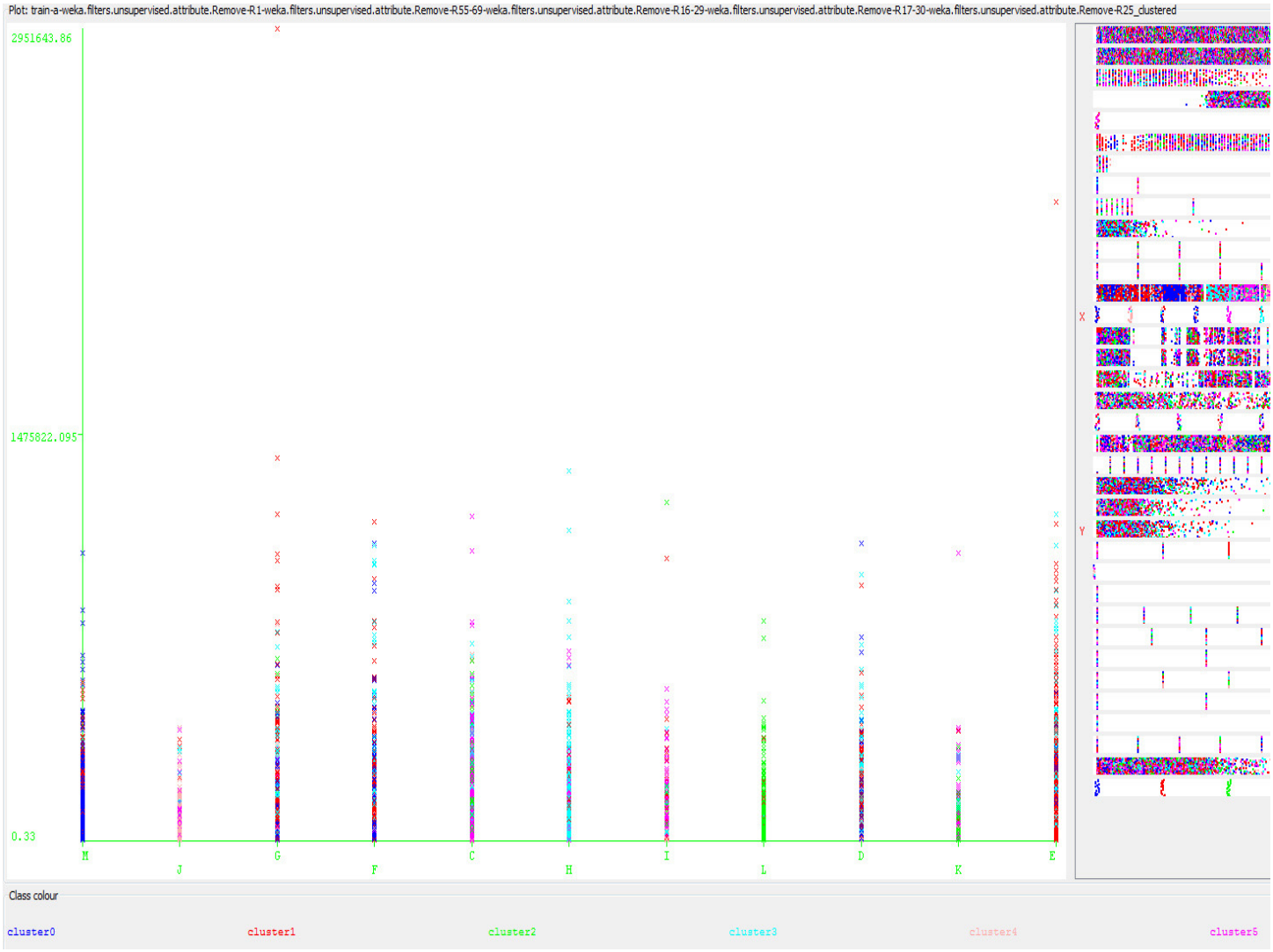
Cluster centroids:

		Cluster#					
Attribute	Full Data	0	1	2	3	4	5
	(99999)	(20664)	(20006)	(12324)	(10136)	(12995)	(23874)
PatientID	39095.379	39151.8946	38904.8598	38732.5894	38423.2419	38986.4689	39738.0351
Record Count	3.6849	3.7813	4.3674	3.2492	3.3132	3.5548	3.483
Interval	135.3515	130.875	123.8021	143.8854	136.2437	142.4723	140.2441
Patient-Sex	F	F	M	F	F	F	M
Age	72.7984	73.0934	67.9205	74.3658	77.0145	74.2653	73.2329
Patient-Race	13.2684	19.9433	23.6808	6.379	10.0261	8.8237	6.1178
Patient-Ethnicity	2.5291	2.256	2.5186	2.539	2.4583	2.9623	2.5633
Patient-Disposition	5.0899	4.5781	4.771	5.1898	6.3434	5.5978	4.9398
Length of Stay	7.8484	7.5372	8.0277	7.4523	12.1483	6.3609	7.1561
Admit-Type	1.296	1.1404	1.3835	1.4537	1.2766	1.3349	1.2629
Admit-Source	5.7777	6.2126	5.2873	5.5117	5.8493	5.6786	5.9732
Hospital-ID	108.1512	47.5461	44.2671	182.9799	112.7204	139.5107	156.5049
Region-ID	C	M	E	L	H	J	C
Principal-Dx-Code	26999.5876	21561.0961	31929.298	24815.8059	25384.4057	42601.6926	20896.3899
Admit-Dx-Code	28319.7758	19333.9182	36233.4207	20620.3332	26570.3829	56366.2754	18917.0322
Principal-PR-Code	6688.8147	7268.0331	6344.1819	6443.0558	6537.4546	6830.4153	6590.3216
Cause-E-Code	E8788	E8788	E8788	E8788	E8788	E8788	E8788
Place-E-Code	E8490	E8490	E8490	E8490	E8490	E8490	E8490
Reimb DRG	221.0615	214.5592	252.3783	250.8639	229.1309	192.3996	197.2376
Reimb MDC	7.0685	7.1997	7.5481	7.8396	7.6265	6.1652	6.4099
AccomCharges	18609.9772	22147.3597	25174.3363	10061.7219	29547.0278	10971.3756	13974.4465
AncilCharges	15277.3074	10236.681	18634.5818	13422.7798	24170.734	11777.0176	15913.6299
TotalCharges	33887.2846	32384.0407	43808.9181	23484.5017	53717.7617	22748.3932	29888.0764
Serv-Class	1.5029	1.4237	1.6619	1.577	1.4965	1.4424	1.4357
Emergency-Dept-Ind	E	E	E	E	E	E	E
Pot Amb	0.0014	0.0016	0.0022	0.0007	0.0005	0.0012	0.0012
Complication-Minor	0.5856	0.5501	0.4957	0.628	0.9607	0.5139	0.5494
Complication-Sever	0.253	0.2026	0.2459	0.2312	0.5975	0.1387	0.23
Trauma-Minor	0.014	0.0126	0.0092	0.0182	0.0164	0.0153	0.0152
Trauma-Severe	0.0379	0.0251	0.0282	0.0509	0.0596	0.0672	0.025
Trauma-Severity	0.0812	0.0572	0.0637	0.1077	0.1192	0.1339	0.0582
Nosocomial Inf	0.1122	0.058	0.0884	0.0592	0.5433	0.0273	0.0695
Severity	3.8595	3.6989	3.7572	4.0857	4.6705	3.6493	3.7377
Cost Weight	2.173	1.9819	2.245	2.302	2.6047	2.0314	2.1054

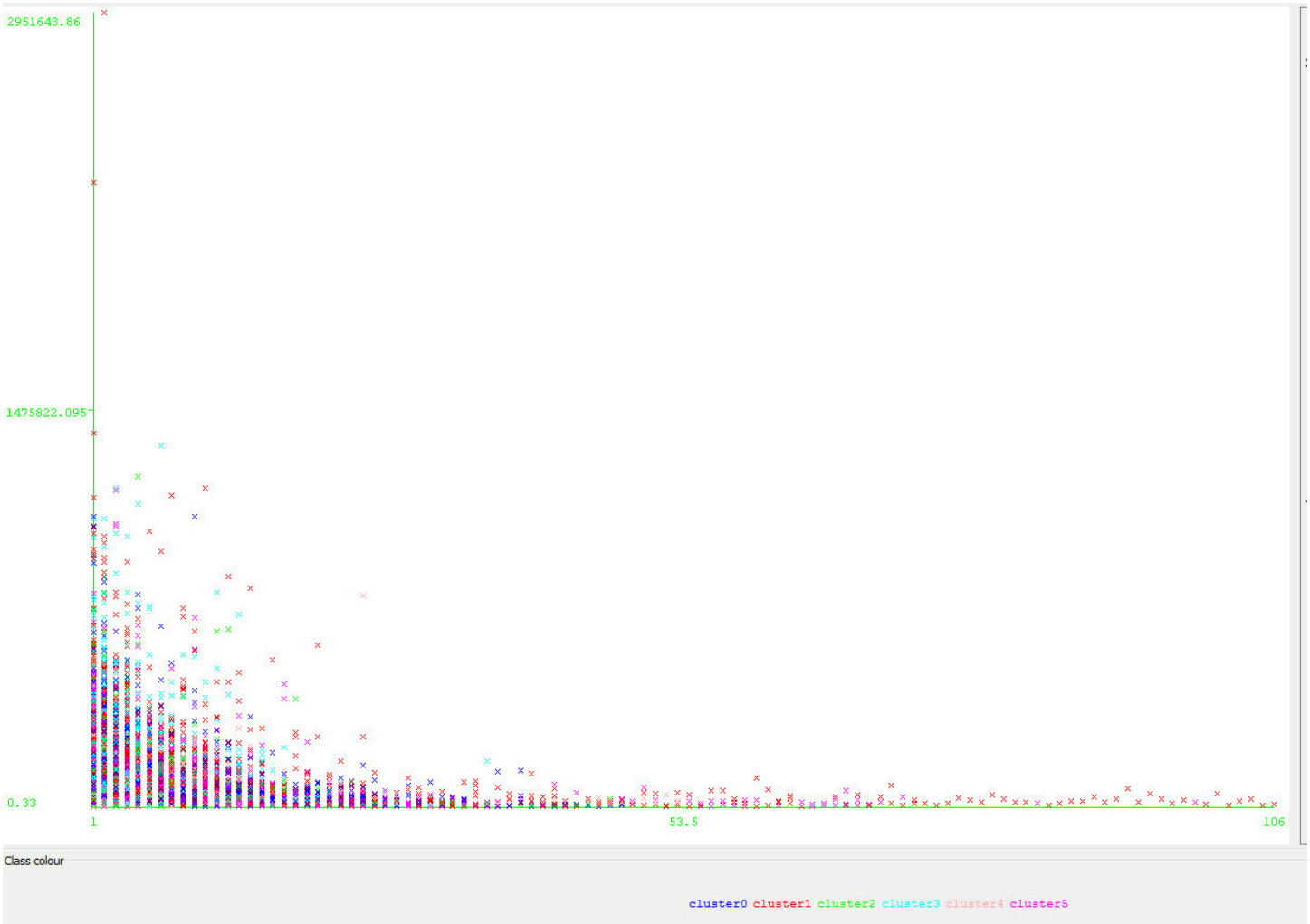
Clustered Instances

0 20664 (21%)
1 20006 (20%)
2 12324 (12%)
3 10136 (10%)
4 12995 (13%)
5 23874 (24%)

Since weka use color to identify clusters, I think to look for color gradients in the visualization of the data would yield more interesting results. By that I mean graphs with well separate colors instead of mixed colors. However those are not the only things I look for.



In the graph above, the x-axis is region ID, and y-axis is the total charges. In the graph the clusters are somewhat well separated. It sort of outlines the cost difference between regions and cost. Some hospitals have high charges that go into millions and some stays low.



In this graph, x-axis is the record count and y axis is the total charges. Eventhough the clusters are so well deparated in this graph, I just think it is interesting that patients with high record count have the lowest charges.

Second run:

I tried the SimpleKmeans again but I chose to ignore certain attributes that are discrete so I would like to compare the continuous attributes . The summary outputs are below(in screenshot pictures):

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -N 6 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10
Relation:     train-a-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.Remove
Instances:    99999
Attributes:   34
              PatientID
              Record Count
              Interval
              Age
              Patient-Disposition
              Length of Stay
              Admit-Type
              Admit-Source
              Hospital-ID
              Region-ID
              AccomCharges
              AncilCharges
              TotalCharges
              Severity
              Cost Weight
Ignored:
              Patient-Sex
              Patient-Race
              Patient-Ethnicity
              Principal-Dx-Code
              Admit-Dx-Code
              Principal-PR-Code
              Cause-E-Code
              Place-E-Code
              Reimb DRG
              Reimb MDC
              Serv-Class
              Emergency-Dept-Ind
              Pot Amb
              Complication-Minor
              Complication-Sever
              Trauma-Minor
              Trauma-Severe
              Trauma-Severity
              Nosocomial Inf
Test mode:    evaluate on training data

=== Model and evaluation on training set ===
```

kMeans
======

Number of iterations: 40
Within cluster sum of squared errors: 65669.35457995304
Missing values globally replaced with mean/mode

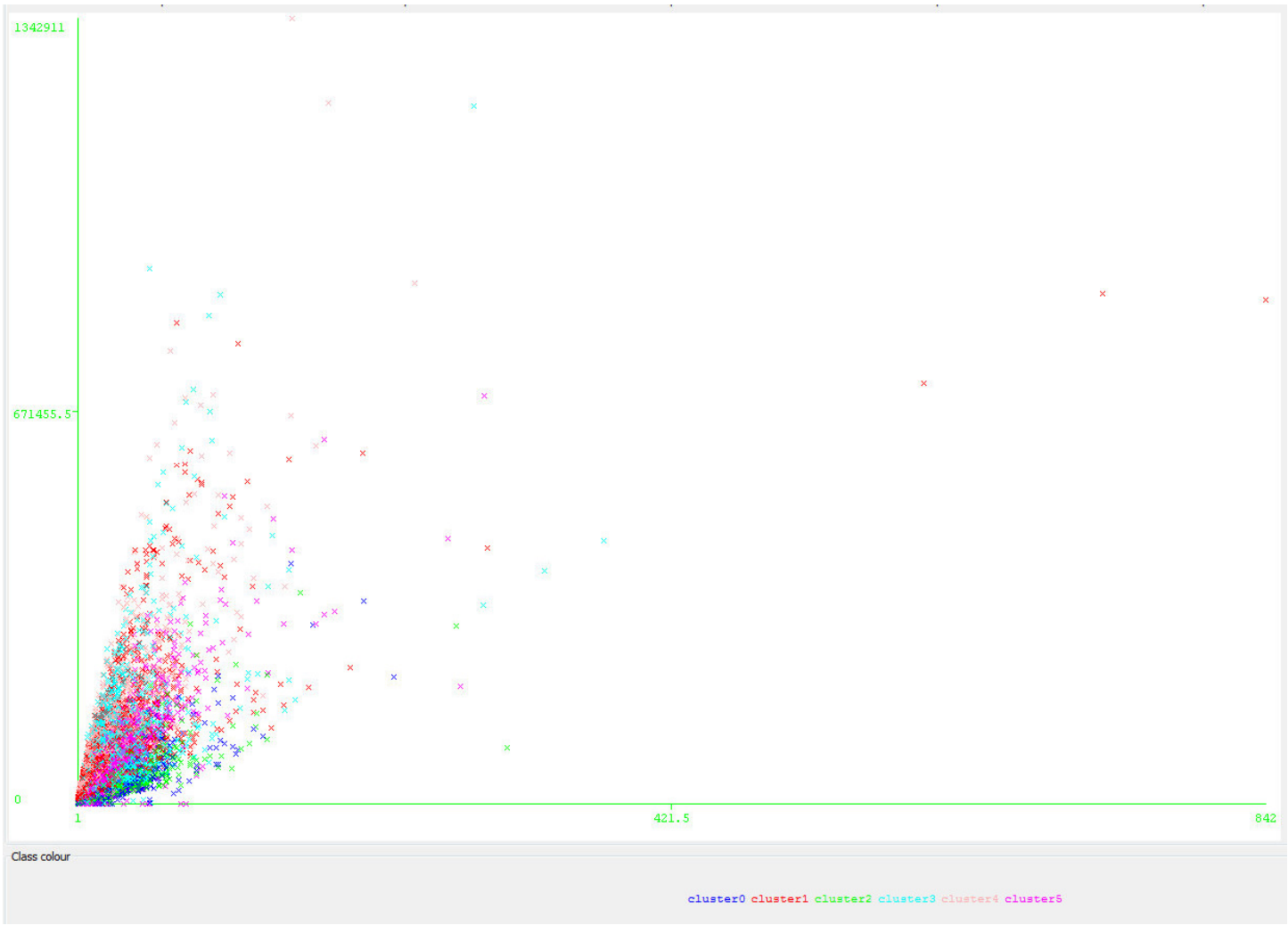
Cluster centroids:

		Cluster#					
Attribute	Full Data	0	1	2	3	4	5
	(99999)	(11211)	(21946)	(14281)	(20493)	(12914)	(19154)
PatientID	39095.379	39252.5314	30152.634	41433.7371	49665.4587	38898.669	36329.8584
Record Count	3.6849	3.4148	3.9211	3.2871	3.8749	4.1237	3.3697
Interval	135.3515	137.476	133.0505	147.9025	135.9523	116.5617	139.412
Age	72.7984	72.8173	71.7619	73.8191	73.1583	69.1694	75.2753
Patient-Disposition	5.0899	5.321	4.4139	5.2993	5.0663	5.4186	5.3766
Length of Stay	7.8484	7.0228	8.0497	7.3042	7.9564	8.2484	8.1215
Admit-Type	1.296	1.7084	1.0894	1.2608	1.1942	1.6273	1.203
Admit-Source	5.7777	4.1975	6.5901	6.3117	6.2671	3.8579	6.1445
Hospital-ID	108.1512	173.5938	37.4719	190.4253	94.869	39.9969	149.6483
Region-ID	C	J	D	L	H	E	C
AccomCharges	18609.9772	6608.7046	23237.169	7562.9394	20285.6677	29266.3666	19591.7107
AncilCharges	15277.3074	14040.3941	12558.9708	11204.8904	14753.3734	21619.648	18436.6385
TotalCharges	33887.2846	20649.0987	35796.1398	18767.8298	35039.0411	50886.0146	38028.3492
Severity	3.8595	3.8867	3.7915	4.036	3.9096	3.7365	3.8195
Cost Weight	2.173	2.2737	2.0411	2.2196	2.1213	2.3426	2.1716

Clustered Instances

0	11211	(11%)
1	21946	(22%)
2	14281	(14%)
3	20493	(20%)
4	12914	(13%)
5	19154	(19%)

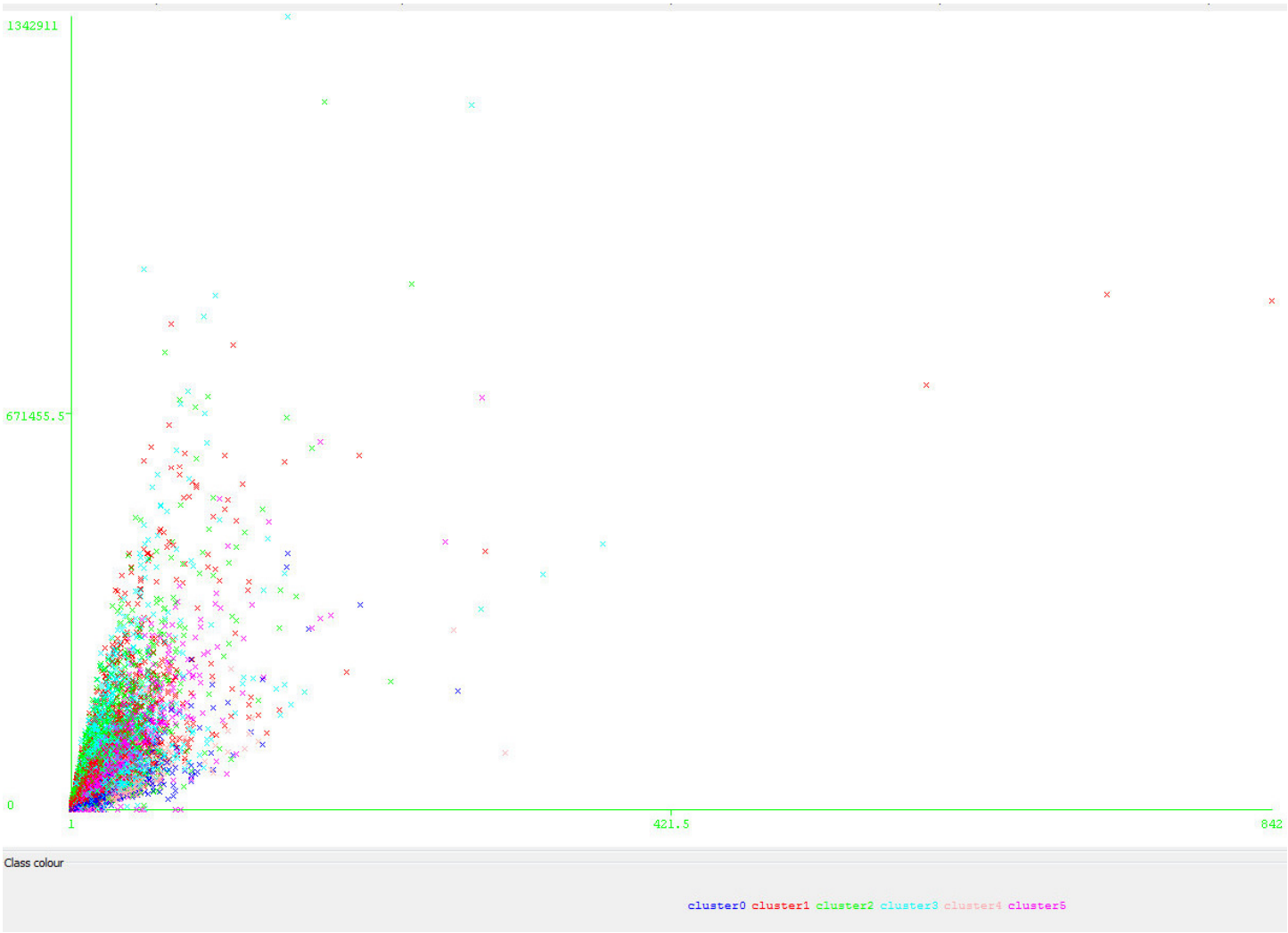
One of the more interesting graphs I see is below:



This graph x axis is length of stay and y axis is accommodation charges. I have noticed the color separation and the semi linear relationship between the two attributes. By the looks of it, the length of stay has different rate of cost for different clusters.

Third Run

I repeated another run, with the same options as the previous run but used Manhattan Distance (emphasis on median instead of mean). The same graph looks as below.



The use of Mahattan made the clustering of colors more separate. Green and light blue clustering involved high cost per length of stay, red and purple indicate medium cost per lengh of stay, and blue indicates low cost per langth of stay.

Fourth Run

I've also tried another clustering algorithm, EM, just to see how the graph would look. I did not particularly understand what EM is about, but did it any way out of curiosity. I used the set itself as a training set just as before. Below is the summary data of the run:

```
=== Run information ===

Scheme:       weka.clusterers.EM -I 100 -N 6 -M 1.0E-6 -S 100
Relation:     train-a-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.R
Instances:    99999
Attributes:   34
              PatientID
              Record Count
              Interval
              Age
              Patient-Disposition
              Length of Stay
              Admit-Type
              Admit-Source
              Hospital-ID
              Region-ID
              AccomCharges
              AncilCharges
              TotalCharges
              Severity
              Cost Weight
Ignored:
              Patient-Sex
              Patient-Race
              Patient-Ethnicity
              Principal-Dx-Code
              Admit-Dx-Code
              Principal-PR-Code
              Cause-E-Code
              Place-E-Code
              Reimb DRG
              Reimb MDC
              Serv-Class
              Emergency-Dept-Ind
              Pot Amb
              Complication-Minor
              Complication-Sever
              Trauma-Minor
              Trauma-Severe
              Trauma-Severity
              Nosocomial Inf
Test mode:    evaluate on training data

=== Model and evaluation on training set ===
```

EM
==

Number of clusters: 6

Attribute	Cluster 0 (0.09)	1 (0.19)	2 (0.1)	3 (0.16)	4 (0.32)	5 (0.14)
PatientID						
mean	38437.2235	39286.3967	39677.9873	39035.0931	39011.5933	39118.9648
std. dev.	22689.691	22624.7623	22547.735	22832.4643	22503.8473	22665.0961
Record Count						
mean	5.6722	2.9604	4.649	4.2533	2.8832	3.5601
std. dev.	5.6065	2.072	8.3141	4.285	2.0678	3.0305
Interval						
mean	103.6646	151.3913	115.8196	125.6266	147.7008	138.3933
std. dev.	124.9832	150.7256	136.3386	133.5459	150.394	144.1434
Age						
mean	70.2062	77.3466	70.4592	70.1372	73.9836	72.8021
std. dev.	15.646	11.6461	15.2866	15.7719	13.8731	14.9377
Patient-Disposition						
mean	3.2862	11.4232	13.6171	2.7137	2.6651	2.7687
std. dev.	3.1308	18.3994	21.1223	2.1231	2.1369	2.098
Length of Stay						
mean	7.4257	10.8968	21.8412	5.6611	4.1197	6.08
std. dev.	5.1831	6.5825	24.4537	4.2259	2.6303	4.6863
Admit-Type						
mean	1.7615	1.0003	1.7159	1	1.4565	1
std. dev.	0.8289	0.0159	0.9363	0.665	0.7722	0.665
Admit-Source						
mean	4.2327	6.9075	4.3953	6.3896	5.392	6.7595
std. dev.	2.7403	0.4536	2.7626	1.7205	2.5665	1.0419
Hospital-ID						
mean	134.0257	154.5728	79.8178	63.2943	159.5246	21.6443
std. dev.	56.871	43.2473	64.016	13.4051	52.3832	11.7117

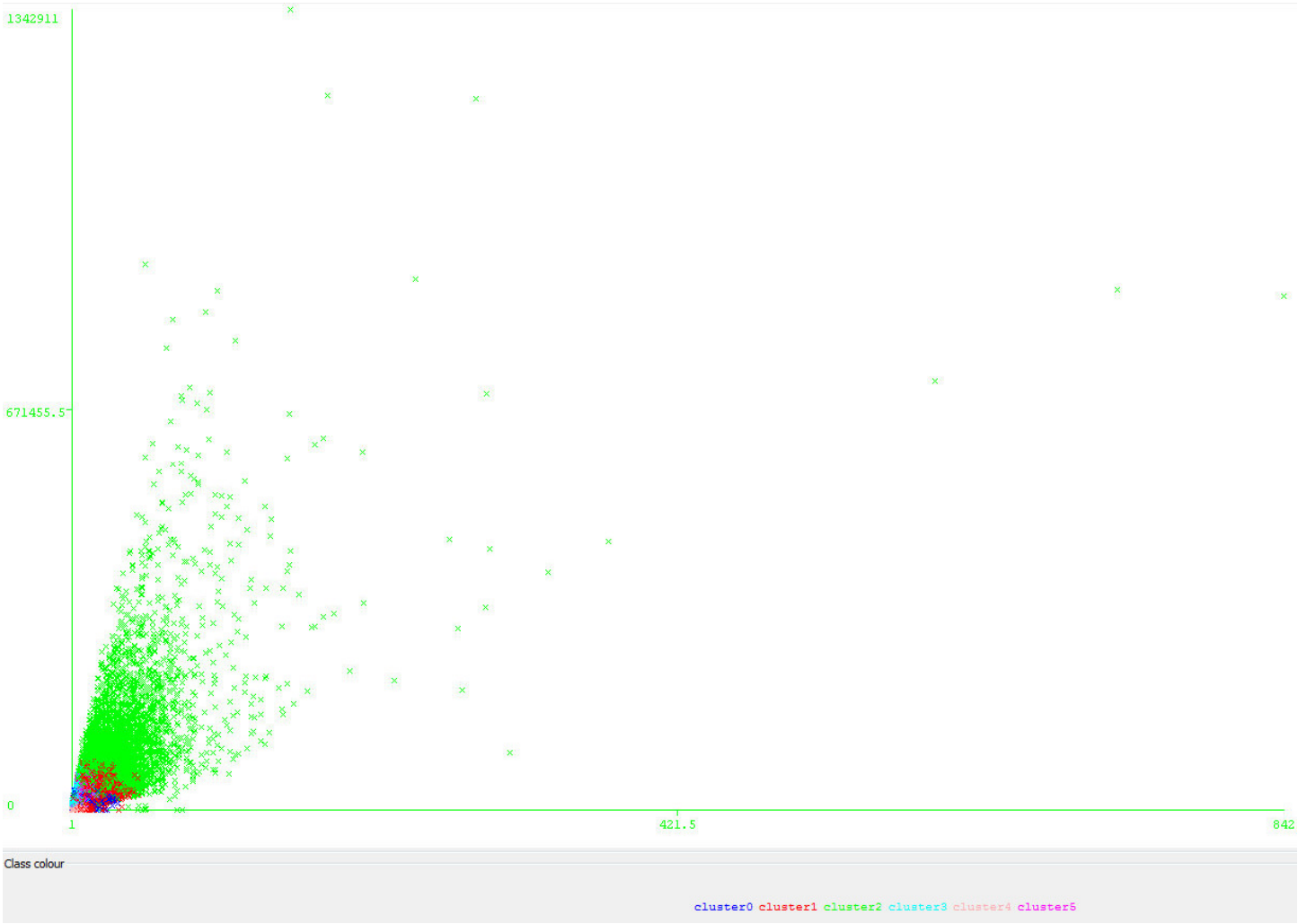
Region-ID						
M	258.2345	190.0242	1475.8442	8789.6981	476.3979	598.8011
J	797.692	805.8949	362.6431	1.6279	6000.142	1
G	168.4467	81.8919	1028.8336	5173.8386	370.8686	860.1205
F	158.5835	9.7459	1060.8737	1810.9726	260.4688	1078.3555
C	1734.3641	6872.4523	1567.2773	2.4423	6108.464	1
H	2697.3609	1663.8726	859.8845	2303.8221	2735.0599	1
I	705.3832	1117.3866	501.0278	3.7302	5361.4722	1
L	547.1595	1114.2844	371.593	29.8805	6108.5632	636.5193
D	109.8619	90.3119	1473.3198	1.3243	462.7836	8276.3985
K	488.4406	772.8167	245.789	2.255	4718.6987	1
E	624.1472	8.799	2333.6096	529.63	755.6966	4304.1175
[total]	8289.6741	12727.4804	11280.6957	18649.2219	33358.6155	15759.3124
AccomCharges						
mean	13158.5986	21435.873	69161.4318	15181.8125	4754.5817	16415.1088
std. dev.	8354.4297	14024.3383	78432.1709	13963.1025	3329.1331	14231.313
AncilCharges						
mean	18519.2247	20477.848	56287.0742	6118.6457	6507.3737	9436.5929
std. dev.	13612.6432	15178.2488	60608.3381	5710.2246	4515.9123	8861.394
TotalCharges						
mean	31677.8233	41913.721	125448.5059	21300.4582	11261.9554	25851.7017
std. dev.	15301.7333	23783.3773	118038.6601	16294.1436	6225.3476	20323.089
Severity						
mean	3.9033	4.6727	5.1347	3.4798	3.4289	3.6288
std. dev.	1.6777	1.6658	1.7524	1.686	1.5829	1.7292
Cost Weight						
mean	2.2874	2.6348	3.5426	1.7986	1.8448	1.8981
std. dev.	1.162	1.194	1.9205	0.9119	0.871	0.9853

Clustered Instances

0	9103	(9%)
1	19837	(20%)
2	9494	(9%)
3	15447	(15%)
4	31719	(32%)
5	14399	(14%)

Log likelihood: -75.86537

The graph of length of stay vs. Accommodation cost is as below:



We can see EM forms completely different cluster compared to KMeans. The conic color spray shape it has a different gradient direction. It seems it falls into certain cluster depends on the radius from the origin, where thevast majority is in the green cluster. As a point gets closer to the origin it becomes red, then blue, then purple.

That concludes my cluster analysis for now. I made some other runs with different options and algorithms but kept the more interesting one. Almost all of the run I did the training on the dataset itself.